

A Concise Information-Theoretic Derivation of the Baum-Welch algorithm

Alireza Nejati, Charles Unsworth

June 27, 2014

Abstract

We derive the Baum-Welch algorithm for hidden Markov models (HMMs) through an information-theoretical approach using cross-entropy instead of the Lagrange multiplier approach which is universal in machine learning literature. The proposed approach provides a more concise derivation of the Baum-Welch method and naturally generalizes to multiple observations.

Introduction

The basic hidden Markov model (HMM)[5] is defined as having a sequence of *hidden* or *latent* states $Q = \{q_t\} = \{q_1, q_2, \dots, q_T\}$ (where t denotes time interval), and each state is statistically independent of all but the state immediately before it and where each state emits some observation o_t with a stationary (non-time-varying) probability density. Formally, the model is defined as:

$$p(O, Q|\lambda) = p(q_1|\lambda) \left[\prod_{t=2}^T p(q_t|q_{t-1}, \lambda) \right] \left[\prod_{t=1}^T p(o_t|q_t, \lambda) \right] \quad (1)$$

Where $\lambda = (\pi, a, b)$ is a set of model parameters. The probability of being in an initial state q_1 is given by the function $p(q_1|\lambda) = \pi(q_1)$. The probability of transitioning from state q_t to state q_{t+1} is given by $p(q_t|q_{t-1}, \lambda) = a(q_t, q_{t-1})$. Finally, the emission density is $p(o_t|q_t) = b(o_t, q_t)$. Given an observation sequence O , it is desirable to find a set of parameters that will maximize the likelihood of producing O . Generally, finding the most optimal set of parameters may be computationally difficult; an approximation is to use the well-known Expectation-Maximization (EM) algorithm [1]. The special case of the EM algorithm applied to hidden Markov models is known as the Baum-Welch algorithm [1][2]. The usual approach to deriving the Baum-Welch is through the use of Lagrange multipliers [3][4]. In this article, we demonstrate that this approach can be improved upon using a method based on cross-entropy that is more concise and lends itself more easily to various HMM generalizations, such as multiple observation sequences.

The Expectation-Maximization Algorithm

For the sake of notational simplicity, we will use $E_Q[\cdot]$ to mean the expected value of the expression inside the brackets over Q given the data and the prior model: (O, λ') . The EM algorithm is as follows. Given an existing set of model parameters λ' and set of observations O , find a new set of model parameters λ such that the following function is maximized:

$$\mathcal{Q}(\lambda, \lambda') = E_Q[\log p(O, Q|\lambda)] \quad (2)$$

By (1), we may rewrite this as:

$$\mathcal{Q}(\lambda, \lambda') = E_Q[\log p(q_1|\lambda)] + \sum_{t=2}^T E_Q[p(q_t|q_{t-1}, \lambda)] + \sum_{t=1}^T E_Q[\log p(o_t|q_t, \lambda)] \quad (3)$$

For each term, we now only use the components of $\lambda = (\pi, a, b)$ that the term depends on and take the expectation over the time-steps of q that are used:

$$\begin{aligned} \mathcal{Q}(\lambda, \lambda') = & E_{q_1}[\log p(q_1|\pi)] + \\ & \sum_{t=2}^T E_{(q_{t-1}, q_t)}[\log p(q_t|q_{t-1}, a)] + \sum_{t=1}^T E_{q_t}[\log p(o_t|q_t, b)] \end{aligned} \quad (4)$$

We note, now, that each term involves optimization of a separate variable and thus the terms can be optimized separately. For the first term in (4), $E_{q_1}[\log p(q_1|\pi)] = -H(p(q_1|O, \lambda'), p(q_1|\pi))$, where H denotes cross entropy. Thus, one must find a π that will minimize the cross entropy between $p(q_1|O, \lambda')$ and $p(q_1|\pi)$. To minimize cross entropy, it suffices to set the distributions to be equal i.e. set π_i to be $P(q_1 = i|O, \lambda')$ for all states i . This value can be computed using the forward-backward algorithm.

The procedure for the second term in (4) is similar. By writing out the expectation explicitly, the following is obtained:

$$\begin{aligned} \sum_{t=2}^T E_{(q_{t-1}, q_t)}[\log p(q_t|q_{t-1}, a)] = \\ \sum_{t=2}^T \sum_{i=1}^N \sum_{j=1}^N p(q_t = j, q_{t-1} = i|O, \lambda') \log p(q_t = j|q_{t-1} = i, a_{ij}) \end{aligned} \quad (5)$$

Noting that the term in front of the logarithm is independent of t , we may rewrite this as follows:

$$= \sum_{i=1}^N \sum_{j=1}^N \log p(q_2 = j|q_1 = i, a_{ij}) \sum_{t=2}^T p(q_t = j, q_{t-1} = i|O, \lambda') \quad (6)$$

Now define the new density α_i such that $\forall j P(q_t = j) = \frac{1}{\gamma_i} \sum_{t=2}^T p(q_t = j, q_{t-1} = i | O, \lambda')$ where γ_i is some normalizing constant to make this a probability density. Now, we may rewrite the previous equation in terms of expectations:

$$= \sum_{i=1}^N E_{\alpha_i} [\log p(q_2 | q_1 = i, a_{ij})] \quad (7)$$

Additionally, define the density a_i to be the density of $P(q_2 = j)$ given a and $q_1 = i$. Thus, the above becomes simply $-\sum_{i=1}^N H(\alpha_i, p(q_2 | a_i))$. Here we are minimizing a sum of independent cross-entropies (each term in the sum is independent since the a_i 's are independent) which is minimized by setting the density a_i to α_i . That is, $a_{ij} = \frac{1}{\gamma_i} \sum_{t=2}^T p(q_t = j, q_{t-1} = i | O, \lambda')$, where $\gamma_i = \sum_{j=1}^N a_{ij}$. These values may be computed, again, using a forward and backward iteration through the chain.

The third term in (4) depends on the probability density used for b , and takes on different forms for discrete, gaussian, or other types of emission distributions. However, in all cases, it is simply equal to the \mathcal{Q} function in the EM algorithm for the mixture model using that specific distribution in question:

$$\mathcal{Q}(b, b') = E_Q \left[\log \prod_{t=1}^T p(o_t | q_t, b) \right] \quad (8)$$

Thus, the problem reduces to performing an expectation-maximization iteration for a mixture model with N mixture components (N = number of possible states for q), where each mixture component has density p , independent of the HMM.

Multiple observations.

The extension of the above algorithm to multiple independent observation sequences [3], then, is straightforward. Consider the case of two observation sequences. The \mathcal{Q} function becomes:

$$\mathcal{Q}(\lambda, \lambda') = E_{q^{(1)} q^{(2)} | (O^{(1)}, O^{(2)}, \lambda')} [\log p(O^{(1)}, q^{(1)}, O^{(2)}, q^{(2)} | \lambda)] \quad (9)$$

Where the expectation, here, is no longer conditional on (O, λ') but is now conditional on $(O^{(1)}, O^{(2)}, \lambda')$. If the observation sequences are independent, this can be written as:

$$\mathcal{Q}(\lambda, \lambda') = E_{q^{(1)} | O^{(1)}, \lambda'} [\log p(O^{(1)}, q^{(1)} | \lambda)] + E_{q^{(2)} | O^{(2)}, \lambda'} [\log p(O^{(2)}, q^{(2)} | \lambda)] \quad (10)$$

We show how the EM algorithm can be simply adapted for this case by considering the optimization of a . Writing out the terms dependent on a , we obtain:

$$\sum_{t=2}^{T^{(1)}} E_{(q_{t-1}^{(1)}, q_t^{(1)})} \left[\log p(q_t^{(1)} | q_{t-1}^{(1)}, a) \right] + \sum_{t=2}^{T^{(2)}} E_{(q_{t-1}^{(2)}, q_t^{(2)})} \left[\log p(q_t^{(2)} | q_{t-1}^{(2)}, a) \right] \quad (11)$$

Or just the following sum:

$$\sum_{k=1}^2 \sum_{t=2}^{T^{(k)}} E_{(q_{t-1}^{(k)}, q_t^{(k)})} \left[\log p(q_t^{(k)} | q_{t-1}^{(k)}, a) \right] \quad (12)$$

Through the same reasoning as employed in the previous section for the sum in the 2nd term, the way to maximize this is simply to set:

$$a_{ij} = \frac{1}{\zeta_i} \sum_{k=1}^2 \sum_{t=2}^{T^{(k)}} p(q_t^{(k)} = j, q_{t-1}^{(k)} = i | O^{(k)}, \lambda') \quad (13)$$

Where $\zeta_i = \sum_{j=1}^N a_{ij}$, as before. A similar reasoning is applied for π , which becomes proportional to $\sum_{k=1}^2 P(q_1^{(k)} = i | O^{(k)}, \lambda')$. The case for more than 2 independent observations is similar. Note that 13 is the same as the formula derived in [3].

Discussion

In this article, we have presented an information-theoretic interpretation of the EM algorithm for hidden Markov models and demonstrated its reduction to simpler, known problems in statistical optimization. We have also demonstrated the use of this method for training HMMs on multiple observation sequences (a situation which arises widely in practice). It is our hope that this new derivation method will be used to provide more fundamental insights into the use of the EM algorithm for various HMM-like models, and perhaps help lead intuition to discovering efficient algorithms for training new custom HMM-like models for various problems.

References

- [1] L. E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, 3(1):8, 1972.
- [2] S Fine, Y Singer, and N Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 62:41–62, 1998.
- [3] Xiaolin Li, Marc Parizeau, and R Plamondon. Training hidden markov models with multiple observations-a combinatorial method. *Pattern Analysis and Machine ...*, 22(4):371–377, 2000.

- [4] Hee-Seon Park and Seong-Whan Lee. A truly 2-D hidden Markov model for off-line handwritten character recognition. *Pattern Recognition*, 31(12):1849–1864, 1998.
- [5] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.